Preprocessing on Web Server Log Data for Web Usage Pattern Discovery

Ketan D. Patel
AMPICS, Ganpat University
Ganpat Vidyanagar
Gujarat, India

Satyen M. Parikh, PhD
AMPICS, Ganpat University
Ganpat Vidyanagar
Gujarat, India

ABSTRACT
World Wide Web has gained popularity because of the fact that it acts as an effective communication medium between business and end users. Company needs to have a web site which satisfies the intended needs of their end users. Users like to revisit a web site which is usable in nature. Web usage patterns of end users must be identified to improve usability on any web site. It is done with analyzing web server log files. Web logs contain noisy, redundant and incomplete data in huge volume which restricts to identify precise usage pattern from it. So, the effective data pre-processing techniques are required. In this paper algorithms are proposed and implemented for pre-processing tasks includes Data Cleaning, User identifications and Session Identification. Pre-processing algorithms are implemented on web log files of two websites and results of these algorithms are useful to study usage pattern of end users.

General Terms
Pre-processing Algorithms, Web Log Files, Usage Mining

Keywords
Preprocessing, Web Server log data, Web Usage Mining, Sessions, Users, Data Cleaning

1. INTRODUCTION
Web sites are great source of information for end users. Companies need to have their web presence to stay in this competitive market. Usable web sites always attract visitors hence study of visitor's usage pattern is necessity to gain useful information. Web usage mining is used to identify usage pattern of end users. Web usage mining which is an application of data mining is used to identify usage behaviors of end users. Web usage mining use log files to discover user behaviors on web site. Web log files capture each and every action of end user he/she performs on the web site. It records the history of user actions performed on the web therefor web server log files are great source of information to study usage patterns.

Typically web server log files contain following information:

- IP address: IP address of visitor who visits the site.
- Page Requested: Requested URL by end user along with Get/Post method and protocol version.
- Referrer: Page from where visitor comes.
- Time stamp: Date and Time zone of visit
- HTTP Code: HTTP status code of returning during visit.
- Bytes served: Number of bytes served during visit
- User agent: Browser and operating system used by visitor during visit.

Study of such logs gives insights about useful information like who visited the site, when, what he/she visited etc. These files are large in volume and contain noisy and irrelevant data in huge volume so pre-processing of such data is required before to apply any further analysis. In this paper algorithms are implemented for data cleaning, user identification and session identification phases of data pre-processing task.

2. RELATED WORK
Research has been attempted by different researchers in the area of pre-processing of server log files. Pre-processing methodologies and few implementations were proposed by different researchers. Suneetha & Dr. Krishnamoorthi identified user behavior by analyzing the NASA web site to find information about potential visitors of the web site and top errors. They proposed overview of data pre-processing steps and identified overall visitors, Hit ratio and total number of unique visitors [1]. Pamutha et al. proposed algorithms for data cleaning and session identification stages. They produce statistical information about user session like Unique IP, Unique pages, total sessions and frequency page visited [2]. Kharwar et al. in their study presented steps and rules to perform data cleaning & session identification and identified and removed robots entries with different approaches [3]. Punjani & Gupta presented the survey based research paper on data pre-processing. The authors proposed steps needed to perform Pre-processing [4]. Dr. Dhavan & Lathwal proposed techniques to perform data pre-processing. The authors proposed steps to perform data cleaning and user identification [5]. Verma & Dr. Kesswani performed two stages of data pre-processing: Data Cleaning and User Identification. Proposed work is focused to clean only .jpg, .gif and .css entries from log file. Authors proposed second approach for user identification phase [6]. Muskan & Dr. Garg proposed algorithm for data storage and cleaning task of data pre-processing. Proposed algorithm cleans robots entries as well as image files and formatting files from server log data [7]. Pushpa & Vidyapriya used web log explorer tool to clean irrelevant entries from log files. Using this tool authors removed spider entries, error code entries, images and other irrelevant entries from server log data to identify user activity [8]. Meghwal & Sharma used web log analyzer tool to find system errors from server log files. Authors identified general statistics of visitors like Hits, Page view, Visitors, Bandwidth and error entries. Findings help web admin to identify broken links from the website [9]. Neelima & Dr. Rodda implemented algorithms for data cleaning, user identification and session identification and visualize results in various forms. They conclude that with pre-processed data further research can be done to analyze usage behaviors of the end users [10].

The above literature study reveals that the web server log study is one of the most parameter to study web usage patterns. Although web server log file is an important entity, the methodologies are not widely available for the accurate data cleansing of web server log files. The web server logs contain the noisy, redundant and incomplete data in huge

volume, which restricts any researcher for accurate study of web usage pattern. Hence, the effective data per-processing techniques and algorithms are highly required.

3. METHODOLOGY

Different algorithms are proposed and implemented in this research to increase accuracy and automation in pre-processing task. All the proposed algorithms are implemented with an intention to provide automation in pre-processing task which result into reduction of human efforts to perform pre-processing.

3.1 Algorithm 1: Data Cleaning

is proposed and implemented to remove irrelevant and missing data from server log files because such data have no relevance in analysis. Algorithm identifies failed error status code and robots/crawler entries from the server log files and eliminates them from further study. Focus of this study is to identify human interaction with website hence crawler visit entries should be removed from log data.

Algorithm 1: Data Cleaning

1. Start
2. For each entry in log file
3. If status_code >=200 and <=299 then
4. If user_agent <> {bot, spider, yandex, crawler} then
5. Save entries in new table.
6. Else
7. Eliminate such entries.
8. End if
9. Else
10. Read next entry
11. End if
12. End

Cleaning Algorithm identifies different crawler visits from user agent field of server log file and eliminates such entries from log files. A web crawler (also called as web spider or web robot) is an automated program or script which browses the World Wide Web [11]. They copy all the pages they visited which are later used by search engines to build indexes. Algorithm also removes http status code entries for http code below 200 and above 299 because entries with such code indicate error or failure task. Cleaning error code entries and crawler visit entries from log files are not sufficient because still web log file contains some irrelevant data in the form of supportive formatting and media files with extensions like .js, .css as well as .gif, .png, .jpg etc. Such entries should be removed just because they are supportive files and are less important for further study. Algorithm 2 is proposed and implemented to eliminate supportive files from the log file.

3.2 Algorithm 2: Data Cleaning (Supportive file)

1. Start
2. For each entry in log file
3. If pageURL.extension = { .js, .css, .gif, .png, .jpg, .svg} then
4. Eliminate such entry
5. Else
6. Save entries in new table
7. End if
8. End

Algorithms are implemented in PHP scripting language. For implementation and testing, web server log data is collected from two web sites named www.ganpatuniversity.ac.in and www.dcs.gnu.ac.in. Data is taken for 20 days (30 Nov to 20 Dec 2016) for GNU site and 30 days (30 Nov to 30 Dec 2016) for DCS site. After collecting the server log data they were converted into .csv format and then imported using PHP script. Algorithm 1 and 2 are applied on the imported data. Results are presented in following tables. Table 1 and 2 represents the cleaning statistics of GNU and DCS web sites respectively. Algorithm identifies and cleans 94.76% of irrelevant data from GNU web log files and 85.26% from DCS log files.

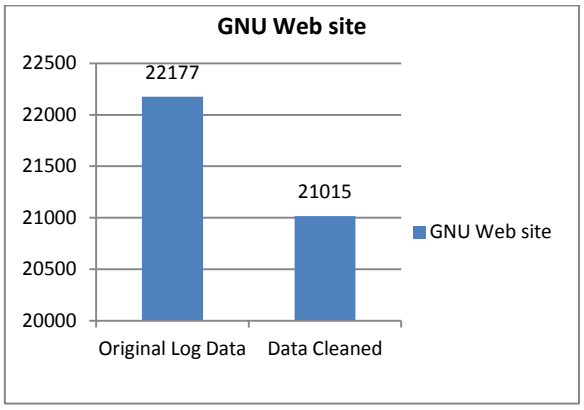
Table 1. Statistics of data cleaning (GNU Site)

Original Log data (GNU Site)	Removed(failed error code , robots and supportive files)	Data Remaining	% of reduction
22177	21015	1162	94.76

Table 2. Statistics of data cleaning (DCS Site)

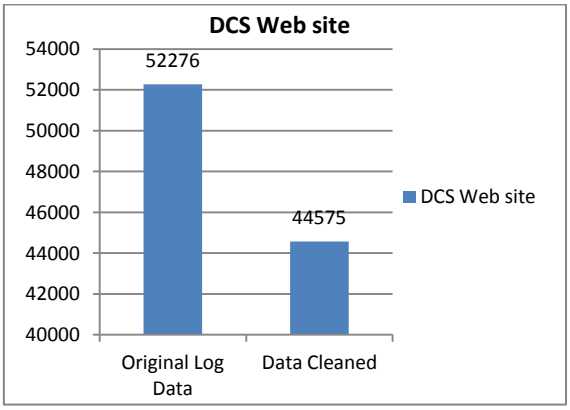
Original Log data (DCS Site)	Removed(failed error code , robots and supportive files)	Data Remaining	% of reduction
52276	44575	7701	85.26

Graph 1 shows the graphical representations of cleaning statistics of GNU site.



Graph 1: Data Cleaning Statistics (GNU web site)

Graph 2 displays the graphical representations of cleaning statistics of DCS site.



Graph 2: Data Cleaning Statistics (DCS web site)

After completion of data cleaning next important task of pre-processing is user action tracking.

Tracking User Action task is mainly focused to identify users and their sessions during their visit to the website. This is used to discover who had visited our pages and what they visited. It helps to identify users and their needs for particular website.

Algorithm 3 is proposed to identify no of visitors of the site. Algorithm identifies unique visitors by matching IP address. If IP address is different means user is considered as new user and if IP address matches with the existing IP list means that user is considered as old user. In case of Proxy server it is bit difficult to differentiate users, hence algorithm assigns new user id when IP address is same but browser or operating system is different.

3.3 Algorithm 3: Unique Visitors Identification

1. Start
2. For each record in log table
3. Repeat for each IP address
4. If IP address is found in IPList and user_agent is same then
5. Assign old userid to that entry
6. Else
7. Assign new userid to that entry.
8. Increment userid.
9. End if
10. End

Table 3 & 4 represents the number of unique visitors identified for DCS & GNU web sites which are 2797 and 479 respectively.

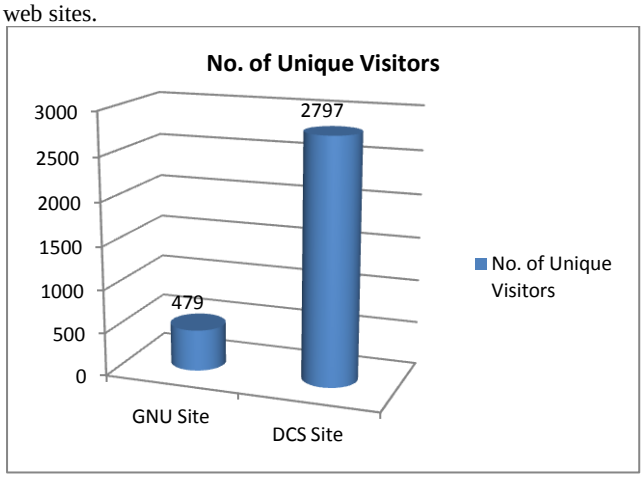
Table 3. No. of Unique Visitors

Web Sites	No. of Unique Visitors (30 Nov to 30 Dec)
DCS	2797

Table 4. No. of Unique Visitors

Web Site	No. of Unique Visitors (30 Nov to 20 Dec)
GNU	479

Graph 3 shows the result of algorithm 3 which is applied on two web sites.



Graph 3: No of Unique Visitors

After identification of visitors, the next important task of pre-processing is identification of user sessions. Identification of user sessions helps the web administrator to know what pages have been visited by particular visitor during his/her visit to a specific period of time. It helps to discover usage patterns of end users.

Algorithm 4 is proposed and implemented to identify user sessions. Algorithm identifies and assign session id to each entry in the log file. It assigns new session id for each user if date of visit is different or time difference is more than 30 minutes (ideal session time) in case of same date visit.

3.4 Algorithm 4: Session Identification

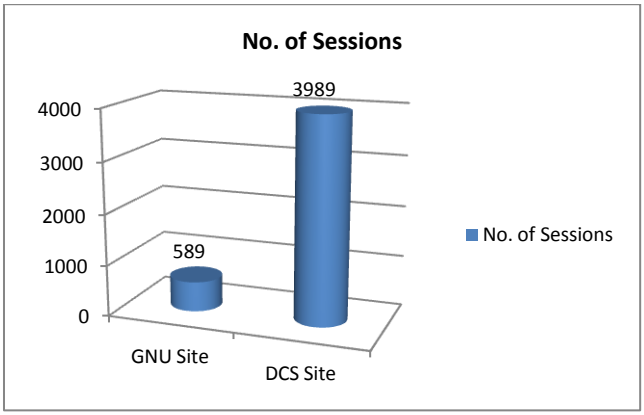
1. Start
2. For each entry in log table
3. If IP address is found in IPList and datetime_difference <= 30 minutes then
4. Assign old sessionid
5. Else
6. Assign the new sessionid
7. Increment sessionid
8. End if
9. End

Table: 5 describe the results of session identification algorithm.

Table 5. No of sessions

Web Sites	No. of Sessions
GNU	589
DCS	3989

Graph 4 visualizes the number of sessions identified in DCS and GNU web sites after applying session identification algorithm.



Graph 4: No of sessions

4. CONCLUSION & FUTURE WORK

To study the usage behaviors of visitors is always an interest of web site owners. Web usage mining fulfills the needs of business owners to identify usage patterns of end users. In this paper we proposed and implemented algorithms to perform pre-processing tasks of web usage mining such as data cleaning, user identification and session identification. Algorithms are implemented on log data collected from two different web sites and then data cleaning is performed effectively. Afterward visitors and sessions are identified from cleaned data. The pre-processed data which is the outcome of

pre-processing task is useful to study interesting patterns of end users. In future pattern analysis and pattern discovery techniques can be applied on the preprocessed data to discover useful usage information. Such findings are helpful to web admin to restructure navigation path of web site in a way which is clear, simple and easy to use.

5. REFERENCES

[1] Suneetha, K. R. and Dr. Krishnamoorthi, R. "Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[2] Pamutha, T., Chimphlee, S., Kimpon, C. and Sanguansat, P. "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns", International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol. 2, No. 2, June 2012.

[3] Kharwar, A., Naik, C. and Desai, N. "A Complete Pre Processing Method for Web Usage Mining", International Journal of Emerging Technology and Advanced Engineering, October 2013.

[4] Punjani, M. and Gupta, V. "A Survey on Data Preprocessing in Web Usage Mining", IOSR Journal of Computer Engineering (IOSR-JCE), 2013.

[5] Dr. Dhawan, S. and Lathwal, M. "Study of Preprocessing Methods in Web Server Logs", International Journal of Advanced Research in Computer Science and Software Engineering, 2013.

[6] Verma, P. and Dr. Keswani, N. "Web Usage mining framework for Data Cleaning and IP address Identification", IJASCSE, 2014.

[7] Muskan. and Dr. Garg, K., "An Efficient Algorithm for Data Cleaning of Web Logs with Spider Navigation Removal", International Journal of Computer Application (2250-1797) Volume 6– No.3, May- June 2016.

[8] Pushpa, V. and Vidyapriya V., "An Efficient Preprocessing Method to Detect User Access Patterns from Weblogs", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.9, September- 2016.

[9] Meghwal, A. and Dr. Sharma A. "Identifying System Errors through Web Server Log Files in Web Log Mining", IJCST Vol. 7, Issue 1, 2016.

[10] Neelima, G. and Dr. Rodda, S. "Predicting user behavior through sessions using the web log mining", International Conference on Advances in Human Machine Interaction, IEEE 2016.

[11] Web crawler, ScienceDaily https://www.sciencedaily.com/terms/web_crawler.htm